Improving Drug Discovery using Hybrid Softcomputing Methods

Horacio Pérez-Sánchez[1*], Gaspar Cano[2] and José García-Rodríguez[2]

[1]Computer Science Department, Catholic University of Murcia (UCAM) E30107 Murcia, Spain E-mail: hperez@ucam.edu

[2]Computing Technology Department, University of Alicante, Ap. 99. E03080. Alicante, Spain. E-mail: {gcano, jgarcia}@dtic.ua.es

[*]corresponding author

Abstract. Virtual Screening (VS) methods can considerably aid clinical research, predicting how ligands interact with drug targets. Most VS methods suppose a unique binding site for the target, but it has been demonstrated that diverse ligands interact with unrelated parts of the target and many VS methods do not take into account this relevant fact. This problem is circumvented by a novel VS methodology named BINDSURF that scans the whole protein surface in order to find new hotspots, where ligands might potentially interact with, and which is implemented in last generation massively parallel GPU hardware, allowing fast processing of large ligand databases. BINDSURF can thus be used in drug discovery, drug design, drug repurposing and therefore helps considerably in clinical research. However, the accuracy of most VS methods and concretely BINDSURF is constrained by limitations in the scoring function that describes biomolecular interactions, and even nowadays these uncertainties are not completely understood. In order to improve accuracy of the scoring functions used in BINDSURF we propose a hybrid novel approach where neural networks (NNET) and support vector machines (SVM) methods are trained with databases of known active (drugs) and inactive compounds, being this information exploited afterwards to improve BINDSURF VS predictions.

Keywords: Neural Networks, Support Vector Machines, Clinical Research, Drug Discovery, Virtual Screening, Parallel Computing.

1 INTRODUCTION

In clinical research, it is crucial to determine the safety and effectiveness of current drugs and to accelerate findings in basic research (discovery of new leads and active compounds) into meaningful health outcomes. Both objectives need to process the large data set of protein structures available in biological databases such as PDB (Berman, et al., 2000) and also derived from genomic data using techniques as homology modeling (Sanchez & Sali, 1998). Screenings in lab and compound optimization are expensive and slow methods (Yao, et al., 2009), but bioinformatics can vastly help clinical research

for the mentioned purposes by providing prediction of the toxicity of drugs and activity in non-tested targets, and by evolving discovered active compounds into drugs for the clinical trials.

This can be achieved thanks to the availability of bioinformatics tools and Virtual Screening (VS) methods that allow testing all required hypothesis before clinical trials. Nevertheless current Virtual Screening (VS) methods, such as docking, fail to make good toxicity and activity predictions since they are constrained by the access to computational resources; even the nowadays fastest VS methods cannot process large biological databases in a reasonable time-frame. Therefore, these constraints impose serious limitations in many areas of translational research.

The use of last generation massively parallel hardware architectures such as Graphics Processing Units (GPUs) can tremendously overcome this problem. The GPU has become increasingly popular in the high performance computing arena, by combining impressive computational power with the demanding requirements of real-time graphics and the lucrative mass-market of the gaming industry (Garland & Kirk, 2010). Scientists have exploited this power in arguably every computational domain, and the GPU has emerged as a key resource in applications where parallelism is the common denominator (Garland et al., 2008). To maintain this momentum, new hardware features have been progressively added by NVIDIA to their range of GPUs, with the Fermi architecture (NVIDIA, 2009) being the most recent milestone in this path. Therefore, GPUs are well suited to overcome the lack of computational resources in VS methods, accelerating the required calculations and allowing the introduction of improvements in the biophysical models not a affordable in the past (Pérez-Sánchez & Wenzel, 2011). We have previously worked in this direction, showing how VS methods can benefit from the use of GPUs (Guerrero, et al., 2011; Sánchez-Linares et al., 2011; Sánchez-Linares et al., 2012a). Moreover, another important lack of VS methods is that they usually take the assumption that the binding site derived from a single crystal structure will be the same for different ligands, while it has been shown that this does not always happen (Brannigan et al., 2010), and thus it is crucial to avoid this very basic supposition. In this work, we present a novel VS methodology called BINDSURF (Sánchez-Linares et al., 2012b) which takes advantage of massively parallel and high arithmetic intensity of GPUs to speed-up the required calculations in low cost and consumption desktop machines, providing new and useful information about targets and thus improving key toxicity and activity predictions. In BINDSURF a large ligand database is screened against the target protein over its whole surface simultaneously. Afterwards, information

obtained about novel potential protein hotspots is used to perform more detailed calculations using particular VS method, but just for a reduced and selected set of ligands.

Other authors have also performed VS studies over whole protein surfaces (Hetényi & van der Spoel, 2002) using different approaches and screening small ligand databases, but as far as we know, none of them have been implemented on GPUs, while BINDSURF has been designed from scratch taken into account the GPU architecture.

However, the accuracy of most VS methods is constrained by limitations in the scoring function that describes biomolecular interactions, and even nowadays these uncertainties are not completely understood. In order to solve this problem we propose a novel hybrid approach where softcomputing methods that includes neural networks (NNET) and support vector machines (SVM) are trained with known active (drugs) and inactive compounds and are later used to improve VS predictions.

The rest of the paper is organized as follows. Section 2 describes the methodology including VS using BINDSURF, NNET and SVM techniques, and molecular properties used in this study. Section 3 presents the experiments carried out to refine the BINDSURF method with the previously mentioned techniques while section 4 discusses the results obtained. In section 5 we present our main conclusions and further work.

2 METHOLODOGY

In this section we describe the methodologies we used for improving the prediction of protein-ligand affinity; a) the Virtual Screening method BINDSURF, and b) two different soft computing techniques are studied; neural networks (NN) and support vector machines (SVM) trained with different molecular properties calculated for known active and inactive compounds selected from standard VS benchmarks. In Figure 1 a flowchart of the methodology is shown; once a protein target (component A) and a compound database (component B) have been chosen, compounds for which no information about affinity against protein target is available (component C) are docked using BINDSURF (component D) and estimated affinities (component E) and 3D poses (component F) are obtained. Using the methods described in this section, we start selecting compounds from the database for which affinity data is available (component G), so that we can calculate relevant descriptors (component H) and train adequately neural networks and support vector machines (component I) so that affinities obtained in component E are post-processed and we finally obtain improved values for the affinities (component J).

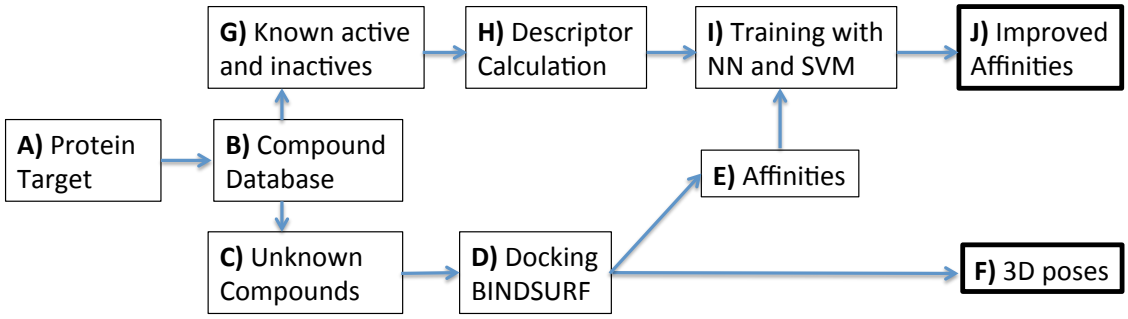

Figure 1. Flowchart of the methodology used for improving the predictive capability of BINDSURF.

2.1 Virtual Screening with BINDSURF

The main idea underlying our VS method BINDSURF is the protein surface screening method, implemented in parallel on GPUs. Essentially, VS methods screen a large database of molecules in order to find which one fit some established criteria (Jorgensen, 2004). In the case of the discovery of new leads, compound optimization, toxicity evaluation and additional stages of the drug discovery process, we screen a large compound database to find a small molecule which interacts in a desired way with one or many different receptors. Among the many available VS methods for this purpose we decided to use protein-ligand docking (Yuriev et al., 2011; Huang & Zou, 2010). These methods try to obtain rapid and accurate predictions of the 3D conformation a ligand adopts when it interacts with a given protein target, and also the strength of this union, in terms of its scoring function value. Docking simulations are typically carried out using a very concrete part of the protein surface in methods like Autodock (Morris et al., 1998), Glide (Friesner et al., 2004) and DOCK (Ewing et al., 2001), to name a few. This region is commonly derived from the position of a particular ligand in the crystal structure, or from the crystal structure of the protein without any ligand. The former can be performed when the protein is co-crystallized with the ligand, but it might happen that no crystal structure of this ligand-protein pair is at disposal. Nevertheless, the main problem is to take the assumption, once the binding site is specified, that many different ligands will interact with the protein in the same region, discarding completely the other areas of the protein.

Given this problem we propose to overcome it by dividing the whole protein surface into defined regions. Next, docking simulations for each ligand are performed simultaneously in all the specified protein spots. Following this approach, new hotspots might be found after the examination of the distribution of scoring function values over the entire protein surface. This information could lead to the discovery of novel binding sites. If we compare this approach with a typical docking simulation performed only in a region of the surface, the main drawback of this approach lies on its increased computational cost. We decided to pursue in this direction and show how this limitation can be overcome thanks to GPU hardware and new algorithmic designs.

In essence, in a docking simulation we calculate the ligand-protein interaction energy for a given starting configuration of the system, which is represented by a scoring function (Wang et al., 2004). In BINDSURF the scoring function calculates electrostatic (ES), Van der Waals (VDW) and hydrogen bond (HBOND) terms.

Furthermore, in docking methods it is normally assumed (Jorgensen, 2004) that the minima of the scoring function, among all ligand-protein conformations, will accurately represent the conformation the system adopts when the ligand binds to the protein. Thus, when the simulation starts, we try to minimize the value of the scoring function by continuously performing random or predefined perturbations of the system, calculating for each step the new value of the scoring function, and accepting it or not following different approaches like the Monte Carlo minimization method (Metropolis et al., 1953) or others. Simulations were always carried out with a total of 500 Monte Carlo steps. For a detailed discussion it is advisable to have a look at our previous BINSURF publication (Sánchez-Linares et al., 2012b).

2.2 Softcomputing Methods

We review the softcomputing methods we will apply to refine the prediction capacities of BINDSURF.

2.2.1 Neural Networks

One of the most dominant application areas of neural networks is non-linear function approximation. The main advantage of neural network modeling is that complex non-linear relationships can be modeled without assumptions about the form of the model. That feature is very useful in the field of drug design and discovery.

In the last years a large number of authors have designed hybrid methods that combined neural networks with other techniques to solve chemistry related problems.

More than two decades ago, the aqueous solubility of organic compounds was studied using neural approaches (Bodor, 1991). In next decade, supervised and unsupervised neural models were employed to model QSAR, predict molecules activities and structure, clustering and many more. (Schneider & Wrede 1998; Peterson, 2000). More recently the problem of drug solubility prediction from structure has been revisited (Jorgensen & Duffy, 2002). The prediction of physico-chemical properties of organic compounds from molecular structure has been extensively studied using hybrid techniques that include neural networks (Taskinen & Yliruusi, 2003; Weisel et al., 2010; Pal & Panja, 2013). Also identification of small-molecule ligands has been improved using neural techniques (Durrant & McCammon, 2010; Durrant and McCammon 2011; Romero Reyes et al., 2013).

There are several types of feed-forward neural networks (NNET), the most widely used being multi-layer networks with sigmoidal activation functions (multi-layer perceptrons) and single layer networks with local activation functions (radial basis function networks). The good approximation capability of neural networks has been widely demonstrated by both practical applications and theoretical research. We decided to use a single-hidden-layer neural network with skip-layer connections in this study (see Figure 2) since it has been clearly demonstrated its impact on the differentiation between active and inactive compounds and other chemical applications (Schneider & Wrede 1998). For such purpose we used the *nnet* function of the R package (Venables & Ripley, 2002).

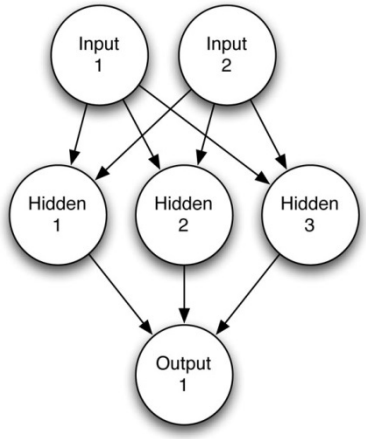

Figure 2. Single Hidden layer Neural Network

2.2.2 Support Vector Machines

Support vector machines (SVM) (Cortes & Vapnic, 1995) are a group of supervised learning methods that can be applied to classification or regression. They represent the decision boundary in terms of a typically small subset of all training examples, called the support vectors. In a short period of time, SVM have found numerous applications in chemistry, such as in drug design (discriminating between ligands and nonligands, inhibitors and noninhibitors, etc.) (Jorissen and Gilson, 2005), drug discovery (Warmuth et al., 2003), quantitative structure-activity relationships (QSAR, where SVM regression is used to predict various physical, chemical, or biological properties) (Kriegl et al., 2005), chemometrics (optimization of chromatographic separation or compound concentration prediction from spectral data as examples), and sensors (for qualitative and quantitative prediction from sensor data), chemical engineering (fault detection and modeling of industrial processes) (Lee et al., 2005). An excellent review of SVM applications in chemistry was published by Ivancicuc (Ivancicuc, 2007).

In our case, we exploit the idea that SVM produce a particular hyperplane in feature space, that separates active from inactive compounds, called the maximum margin hyperplane (see Figure 3).

Most used kernels within SVM include: linear (dot), Polynomial, Neural (sigmoid,Tanh), Anova, Fourier, Spline, B Spline, Additive, Tensor and Gaussian Radial Basis or Exponential Radial Basis.

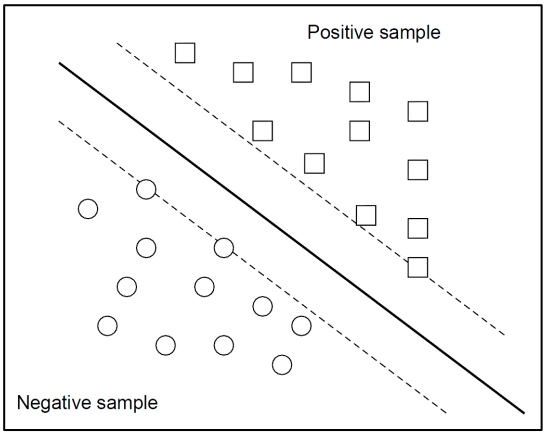

Figure 3. Support Vector Machines margin hyperplanes

2.3 Ligand databases and molecular properties

We carried out our study applying the methods described in sections 2.2.1 and 2.2.2 and using different sets of molecules that are known to be active or inactive. We employed standard VS benchmark tests, such as the Directory of Useful Decoys (DUD) (Huang et al., 2006), where VS methods check how efficient they are in differentiating ligands that are known to bind to a given target, from non-binders or decoys. Input data for each molecule of each set contains information about its molecular structure and whether it is active or not. We focused on three diverse DUD datasets (details are shown in Table 1) that cover kinases, nuclear hormone receptors and other enzymes such as TK, which corresponds to thymidine kinase (from PDB 1KIM), MR, which corresponds to mineralocorticoid receptor (from PDB 2AA2), and GPB, which corresponds to the enzyme glycogen phosphorylase (from PDB 1A8I).

Protein	PDB Code	Resolution (Å)	n_o of ligands	n_o of decoys
GPB	1a8i	1.8	52	1851
MR	2aa2	1.9	15	535
TK	1kim	2.1	22	785

Table 1. Number of active (ligands) and inactive compounds (decoys) for each of the ligand datasets used in this study and obtained from DUD.

Next, using the ChemoPy package (Cao et al., 2013) we calculated for all ligands of the TK, MR and GPB sets a diverse of molecular properties derived from the set of constitutional, CPSA (charged partial surface area) and fragment/fingerprint-based descriptors, as described in Table 2. Constitutional properties depend on very simple descriptors of the molecule that can be easily calculated just counting the number of molecular elements such as atoms, types of atoms, bonds, rings, etc. These descriptors should be able to differentiate very dissimilar molecules, but might have problem for separating closely related isomers. CPSA descriptors take into account finer details of molecular structure, so they might be able to separate similar molecules, but might also have also difficulties for separating isomers. Lastly, fragment and fingerprint-based descriptors take into account the presence of an exact structure (not a substructure) with limited specified attachment points. These descriptors are more difficult to calculate. In generating the fingerprints, the program assigns an initial code to each atom. The initial atom code is derived from the number of connections to the atom, the element type, atomic charge, and atomic mass. This corresponds to an ECFP with a neighborhood size of zero. These atom codes are then updated in an

iterative manner to reflect the codes of each atoms neighbors. In the next iteration, a hashing scheme is employed to incorporate information from each atoms immediate neighbors. Each atoms new code now describes a molecular structure with a neighborhood size of one. This process is carried out for all atoms in the molecule. When the desired neighborhood size is reached, the process is complete and the set of all features is returned as the fingerprint. For the ECFPs employed in this paper, neighborhood sizes of two, four and six (ECFP 2, ECFP 4, ECFP 6) were used to generate the fingerprints. The resulting ECFPs can represent a much larger set of features than other fingerprints and contain a significant number of different structural units crucial for the molecular comparison, among the compounds.

CONSTITUTIONAL DESCRIPTORS	
Natom	Number of atoms
MolWe	Molecular Weight
NRing	Number of rings
NArRg	Numer of aromatic rings
NRotB	Number of rotatable bonds
NHDon	Number of H-bond donors
NHAcc	Number of H-bond acceptors
CPSA DESCRIPTORS	
Msurf	Molecular surface area
Mpola	Molecular polar surface area
Msolu	Molecular solubility
AlogP	Partition coefficient
FRAGMENT/FINGERPRINT-BASED DESCRIPTORS	
ECP2, ECP4, ECP6	Extended-connectivity fingerprints (ECFP)
EstCt	Estate counts
AlCnt	AlogP2 Estate counts
EstKy	Estate keys
MDLPK	MDL public keys

Table 2. Molecular descriptors used in this study.

3. RESULTS AND DISCUSSION

A set of experiments has been carried out in order to test the validity of our initial hypothesis combining and refining BINDSURF results with the proposed softcomputing methods.

3.1 Virtual Screening with BINDSURF

After BINDSURF calculations (depicted by components A, B, C and D in Figure 1), results for three different DUD datasets (depicted by components E and F in Figure 1) are shown in the ROC curves of

Figure 3. Given the results obtained for the DUD datasets TK, MR and GPB, and characterized by the value of the area under the curve (AUC) for each ROC curve, it could be said that, on average, BINDSURF performs similarly well than other docking methods such as DOCK (Shoichet et al., 1992), ICM (Abagyan et al., 1994) and GLIDE (Friesner et al., 2004) as reported for these datasets (Cross et al., 2009) and shown in Table 3.

Dataset	BINDSURF	DOCK	ICM	GLIDE
TK	0.700	0.521	0.723	0.681
MR	0.695	0.554	0.789	0.856
GPB	0.675	0.454	0.462	0.823

Table 3. Obtained values for AUC of the ROC curves for the docking programs BINDSURF, DOCK, ICM and GLIDE when processing DUD datasets TK, MR and GPB.

Nevertheless, it is clear that there is still room for improvement in the scoring function that BINDSURF uses, and on its energy optimization method (Monte Carlo), since both affect directly to the effectiveness of the direct prediction of binding poses.

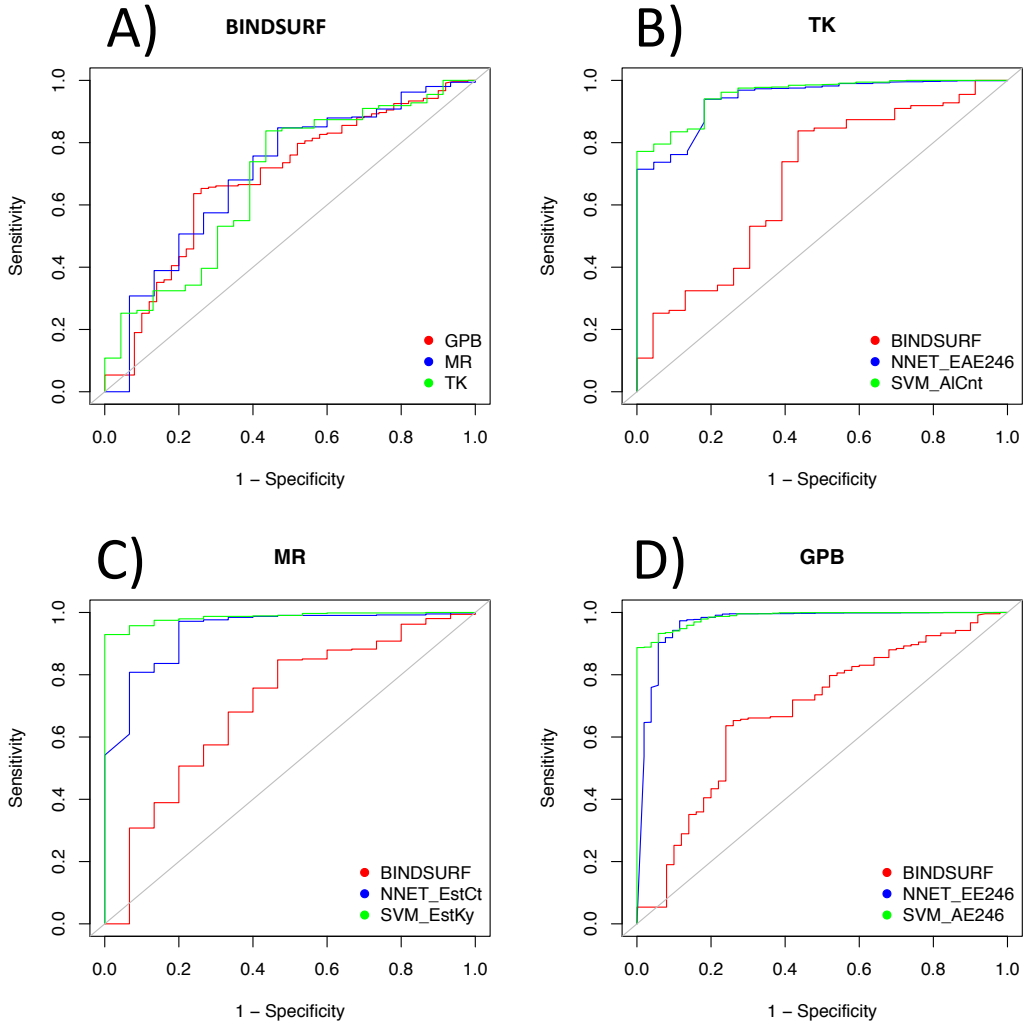

Figure 4. From up-left to down right ROC plots; A) obtained using BINDSURF for the targets of the DUD data set GPB (red), MR (blue) and TK (green), B) obtained for data set TK using BINDSURF (red), the property that yields top AUC value (EAE246) using NNET, and the property that yields top AUC value (AlCnt) using SVM, C) obtained for data set MR using BINDSURF (red), the property that yields top AUC value (EstCt) using NNET, and the property that yields top AUC value (EstKy) using SVM, and D) obtained for data set GPB using BINDSURF (red), the property that yields top AUC value (EE246) using NNET, and the property that yields top AUC value (AE246) using SVM. In all cases diagonal line indicates random performance.

3.2 Activity prediction using Softcomputing methods

NNET and SVM were trained with the previously described DUD datasets TK, MR and GPB (depicted by component G in Figure 1). Molecular properties described in Table 2 were calculated for each molecule (depicted by components H in Figure 1) as described in the methods section. A *k* folds cross-validation technique with *k*=5 was employed for neural networks and SVM experiments (depicted by component I in Figure 1).

3.2.1 Neural Networks

A set of experiments was carried out in order to find the most optimal feedforward neural network architecture for the classification problem proposed. Different numbers of neurons for the hidden layer were tested with the different descriptors and datasets previously described. Best results were obtained with 3 neurons per layer, c(3), for most of the properties and datasets tested. Increasing the number of neurons did not improve the results, as shown in Figure 5.

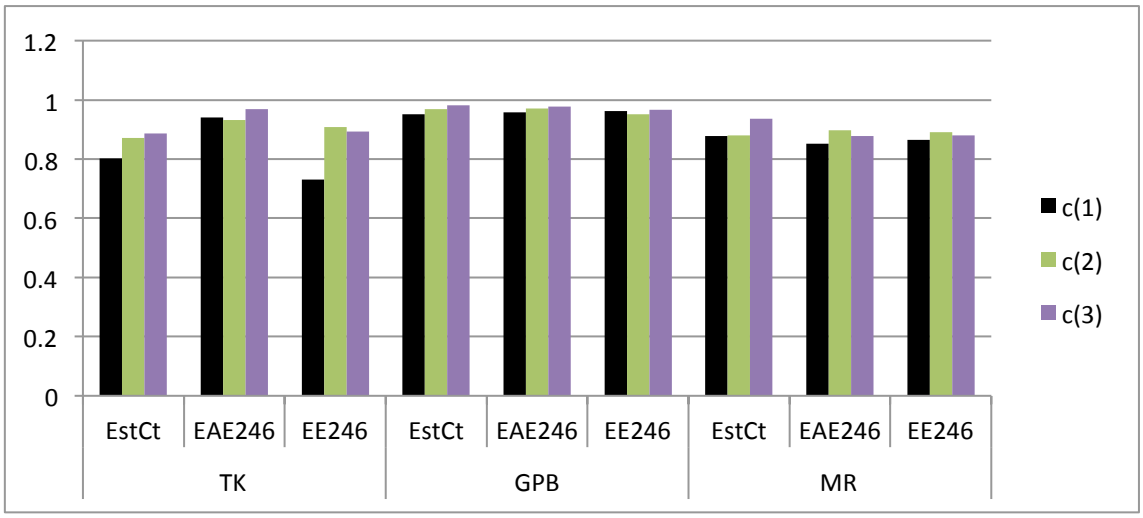

Figure 5. AUC values of the ROC curves for some of the properties calculated in this study for the different datasets TK, GPB and MR, and for different numbers of neurons per layer, c(1), c(2) and c(3) used in NNET.

After having chosen 3 neurons per layer for the NNET architecture, we can observe in Figure 6 results obtained for all properties and different datasets, previously described.

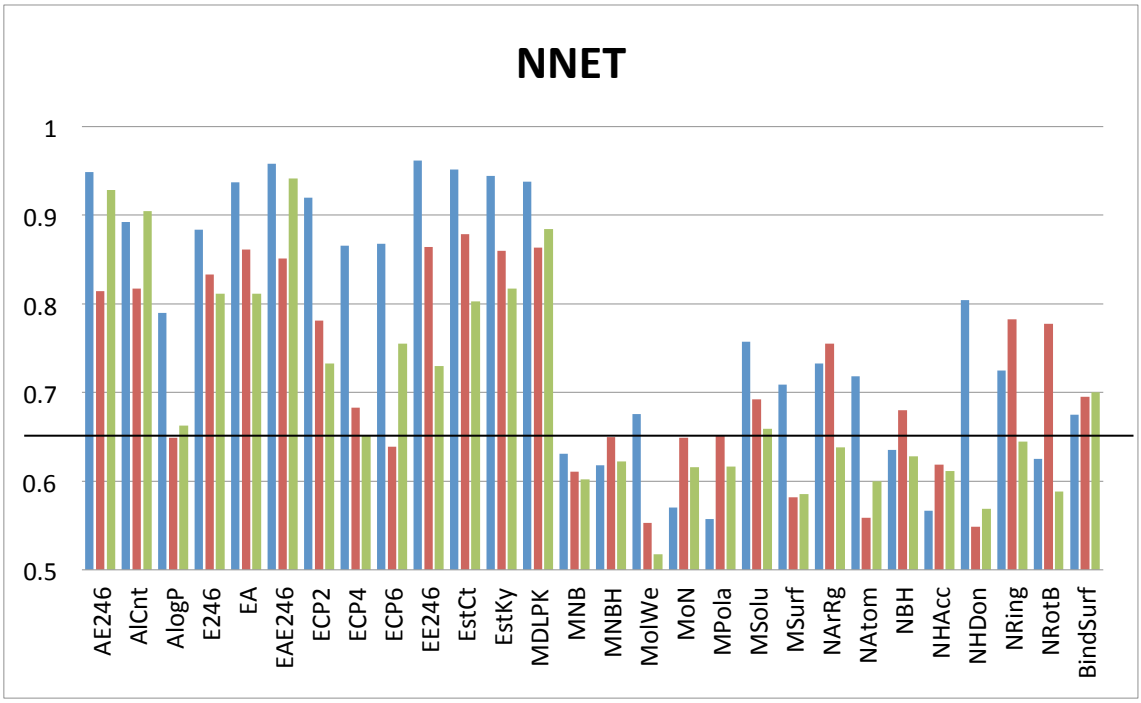

Figure 6. AUC values of the ROC curves obtained using NNET as described in section 2.2.1 for each property of Table 2 of the three different datasets GPB (blue), MR (red) and TK (yellow). Baseline for AUC=0.65 is also shown. The resulting AUC values for the combined properties described in Table 5 are also reported, and also AUC values obtained by BINDSURF as depicted in Figure 4A.

3.2.2 Support Vector Machines

A set of experiments with different kernels was carried out in order to find the option with the best discrimination capacity between active and non-active compounds for each descriptor. More specifically, linear, polynomial, sigmoidal and radial kernels were tested with all the descriptors and datasets, and best results were obtained with radial kernel. The results obtained for AUC values of the most relevant properties for different datasets and with different kernels are reported in Figure 7.

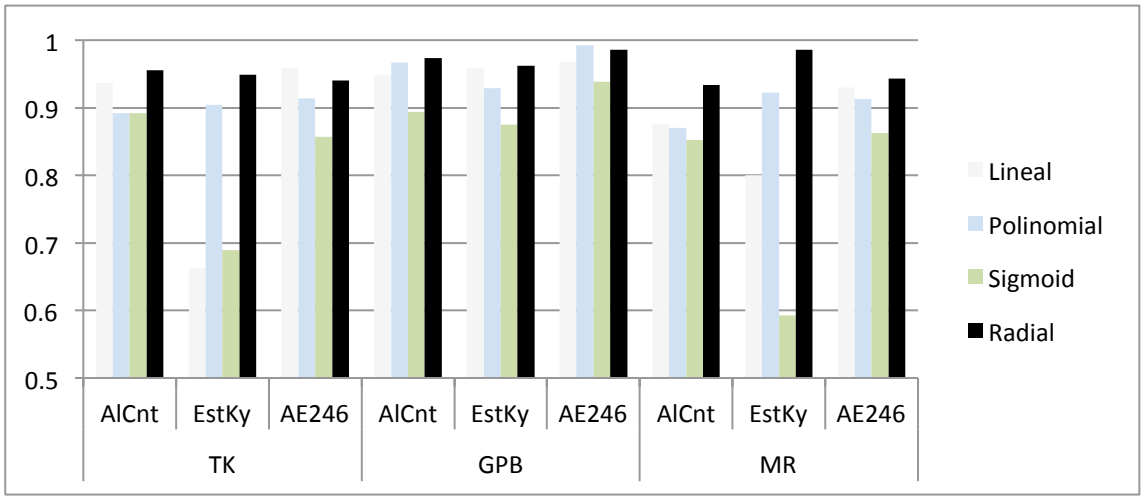

Figure 7. AUC values of the ROC curves of some relevant properties obtained using SVM for different kernels.

We can observe in Figure 8 results obtained with SVM and radial kernel for all properties and different datasets, previously described.

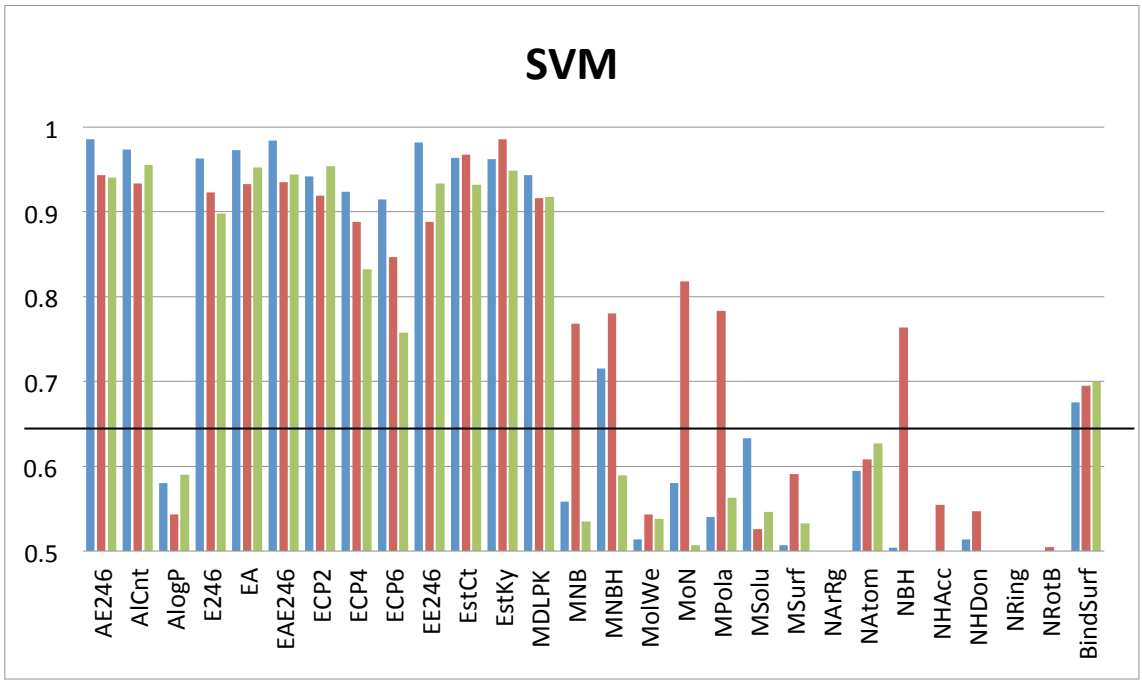

Figure 8. AUC values of the ROC curves obtained using SVM as described in section 2.2.1 for each property of Table 2 of the three different datasets GPB (blue), MR (red) and TK (yellow). Baseline for AUC=0.65 is also shown. The resulting AUC values for the combined properties described in Table 5 are also reported, and also AUC values obtained by BINDSURF as depicted in Figure 4A.

4. DISCUSSION

BINDSURF predictions reach AUC values around 0.7 as reported in Fig. 4 and Table 3. From the other side, AUC values reported by both NNET and SVM depend clearly on the considered molecular property, and to a lesser extent, on the molecular dataset studied (GPB, MR, TK). The reason for the latter might be that main active compounds of these sets have similar structures (as shown in Figure 9) consisting in small molecules with two or four rings, and also because they establish similar interactions with the protein, mainly based on hydrogen bond networks.

We propose a threshold value of 0.65 for AUC in order to discriminate which properties are useful for active/inactive prediction. Properties that yield simultaneously AUC values higher than this threshold for all sets using both NNET and SVM are; AlCnt, E246, ECP2 and MDLPK, while properties that yield AUC values lower than threshold are mostly AlogP, MolWe, MPola, MSolu, MSurf, NArgRg, Natom, NHacc, NHDon, NRing, and NRotB. So it seems clear that the best option for discriminating among active and inactive compounds in these datasets is to use fingerprint-based descriptors and to avoid the use of constitutional and CPSA descriptors. This is reasonable since fingerprint descriptors take into account more details about the structure of molecules, being able to efficiently discriminate with more accuracy between active compounds and their decoys.

Next, we studied whether combination of properties could lead to improvements on the predictive capability of these softcomputing methods. Therefore we combined properties that yielded the lowest AUC values, constitutional descriptors, and the properties that yielded the highest AUC values, so fingerprint based descriptors. Combinations used are described in Table 5 and AUC values obtained are reported in Figures 6 and 8. In the case of combinations of constitutional descriptors, there is no clear improvement for either NNET or SVM, while for fingerprint combinations, average AUC values for the three datasets improve slightly.

TK		MR		GPB	
NNET_EE246	0.94	NNET_EstCt	0.87	NNET_EAE246	0.96
SVM_AE246	0.95	SVM_EstKy	0.98	SVM_AlCnt	0.98
BINDSURF	0.70	BINDSURF	0.70	BINDSURF	0.68

Table 4. Top obtained values for AUC of the ROC curves when processing DUD datasets TK, MR and GPB for BINDSURF, NNET and SVM. For each dataset, the property that yields that top value of AUC for both NNET and SVM is specified.

Finally, top obtained AUC values for datasets GPB, MR and TK correspond to properties EE246 (0.96), EstCt (0.87) and EAE246 (0.94) when using NNET, and AE246 (0.98), EstKy (0.98)and AlCnt (0.95) when using SVM. Obtained ROC curves for the mentioned top properties can be seen in Figure 4. Therefore, if we compare NN and SVM obtained results with the previous ones obtained by BINDSURF, as reported in Figures 4B, 4C and 4D and Table 4, it is clear that predictive capability increases when using the presented methodology with the application of softcomputing methods as depicted in Figure 1.

COMBINATIONS OF CONSTITUTIONAL DESCRIPTORS	
MNBH	Molecular polar surface area (MPola)+ Number of rotatable bonds (NRotB) + Number of H-Bond acceptors (NHAcc)
MNB	Molecular polar surface area (MPola) + Number of rotatable bonds (NRotB)
NBH	Number of rotatable bonds (NRotB) + Number of H-Bond acceptors (NHAcc)
MoN	Molecular polar surface area (MPola) + Number of H-Bond acceptors (NHAcc)
COMBINATIONS OF FRAGMENT/FINGERPRINT-BASED DESCRIPTORS	
EAE246	Estate counts (EstCt) + AlogP2 Estate counts (AlCnt) + Extended-connectivity fingerprints (ECFP)
EA	Estate counts (EstCt) + AlogP2 Estate counts (AlCnt)
AE246	AlogP2 Estate counts (AlCnt) + Extended-connectivity fingerprints (ECFP)
EE246	Estate counts (EstCt) + Extended-connectivity fingerprints (ECFP)

Table 5. Combinations of molecular descriptors used in this study.

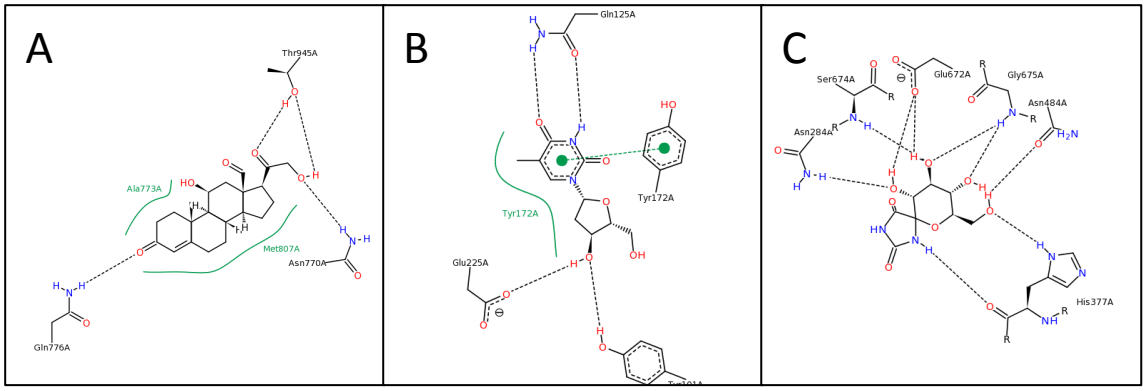

Figure 9. Depiction of the molecular structure and protein-ligand interactions established by main active compounds from A) MR, B) TK, and C) GPB.

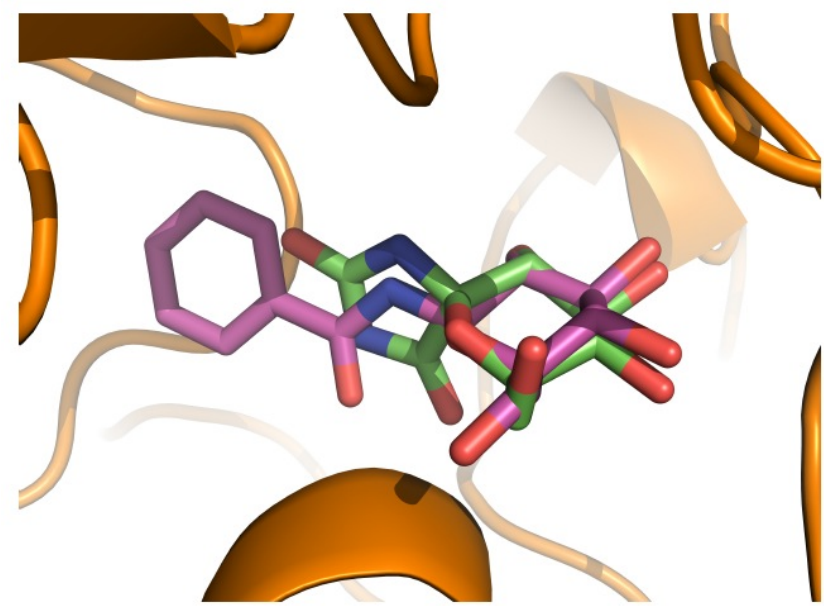

Figure 10, Depiction of the binding mode found by BINDSURF for ligand number 17 of the DUD data set of GPB (PDB ID: 1A8I) with pink skeleton and crystallographic pose for Beta-D-Glucopyranose Spirohydantoin, with green skeleton.

Consequently, and taking into account information obtained by softcomputing methods, we can post-process docking results obtained by the scoring function of BINDSURF (as depicted in the flowchart in Figure 1) and neglect resulting compounds that are predicted as inactive. Then we can sort them by the final affinity value predicted by the BINDSURF scoring function for such cases and study visually the top ones. As an example we can observe in Figure 10 the good agreement in the comparison between the top predicted compound from the DUD database for GPB and the crystallographic pose. In this case main stabilizing interactions are due to a hydrogen-bond network where the intervening nitrogen and oxygen atoms of the predicted compound fall closely to the same atoms of the crystallographic pose.

5. CONCLUSIONS

In this work we have shown how the predictive capability of the VS method BINDSURF can be improved applying softcomputing methods such as neural networks and support vector machines when using only a small set of representative chemical properties. We have also studied which of these properties are the most representative, and we have finally obtained that topological properties can efficiently discriminate between active and non-active compounds for the datasets studied. However, it

must be mentioned that softcomputing approaches can only be used when there is data available for active and non-active compounds for a given protein. For further studies we consider it would be of high interest to train softcomputing methods with a diverse range of absolute affinity data for known compounds and to check whether prediction accuracy still increases with respect to the methodology presented on this work.

Given the improvements shown in the obtained results, we conclude that this methodology can be used to improve drug discovery, drug design, repurposing and therefore aid considerably in biomedical research. In the next steps we want to substitute the Monte Carlo minimization algorithm already present in BINDSURF for more efficient optimization alternatives, such as the Ant Colony optimization method, which we have already efficiently implemented on GPU (Cecilia et al., 2011) and implement also full ligand and receptor flexibility.

ACKNOWLEDGEMENTS

We thank the Catholic University of Murcia (UCAM) under grant PMAFI/26/12. This work was partially supported by the computing facilities of Extremadura Research Centre for Advanced Technologies (CETA-CIEMAT), funded by the European Regional Development Fund (ERDF). CETA-CIEMAT belongs to CIEMAT and the Government of Spain. The authors also thankfully acknowledge the computer resources and the technical support provided by the Plataforma Andaluza de Bioinformática of the University of Málaga. We would also like to thank Dr. Andrés Bueno-Crespo for fruitful discussions.

REFERENCES

Abagyan, R., Totrov, M., Kuznetsov, D. (1994). ICM - a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. Journal of Computational Chemistry 15, 488–506.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res 28, 235-242

Brannigan, G., LeBard, D.N., H_enin, J., Eckenho_, R.G., Klein, M.L. (2010). Multiple binding sites for the general anesthetic isourane identified in the nicotinic acetylcholine receptor transmembrane domain. Proc Natl Acad Sci USA 107(32), 14122-14127

Bodor, N., Harget , A.,Huang, M.J. (1991) Neural network studies. 1. Estimation of the aqueous solubility of organic compounds. Am. Chem. Soc, 113 (25), pp 9480–9483

Cao, D.-S., Xu, Q.-S., Hu, Q.-N. & Liang, Y.-Z. ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics* (2013). doi:10.1093/bioinformatics/btt105

Cecilia, J.M., García, J.M., Ujaldon, M., Nisbet, A., Amos, M. (2011). Parallelization strategies for ant colony optimisation on gpus. In: In 14th Int. Workshop on Nature Inspired Distributed Computing -NIDISC11- (in conjunction with IPDPS 2011), IEEE, 339-346

Cortes, C., Vapnik, V. (1995). Support vector networks. Machine Learning 20, 273-297.

Cross, J.B., Thompson, D.C., Rai, B.K., Baber, J.C., Fan, K.Y., Hu, Y., Humblet, C. (2009). Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. J Chem Inf Model 49(6), 1455-1474

Durrant, J. D., McCammon, J. A. (2010). NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein-Ligand Complexes. J. Chem. Inf. Model, 50 (10), 1865–1871.

Durrant J. D., McCammon J. A. NNScore 2.0. (2011). A Neural-Network Receptor–Ligand Scoring Function. J J Chem Inf Model, 28; 51(11): 2897–2903.

Ewing, T.J.A., Makino, S., Skillman, A.G., Kuntz, I.D.: Dock 4.0. (2001). Search strategies for automated molecular docking of exible molecule databases. J Comput-Aided Mol Des 15(5), 411-428

Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelley, M., Perry, J.K., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. J Med Chem 47(7), 1739-1749

Garland, M., Kirk, D.B. (2010). Understanding throughput-oriented architectures. Commun ACM 53, 58-66

Garland, M., Le Grand, S., Nickolls, J., Anderson, J., Hardwick, J., Morton, S., Phillips, E., Zhang, Y., Volkov, V. (2008). Parallel computing experiences with cuda. IEEE Micro 28, 13-27

Guerrero, G., Pérez-Sánchez, H., Wenzel, W., Cecilia, J.M., García, J.M. (2011). Effective parallelization of non-bonded interactions kernel for virtual screening on gpus. In: 5th International Conference on

Practical Applications of Computational Biology; Bioinformatics (PACBB 2011). Volume 93. Springer Berlin / Heidelberg, 63-69

Hetényi, C., van der Spoel, D. (2002). Efficient docking of peptides to proteins without prior knowledge of the binding site. Protein Sci 11(7), 1729-1737

Hinselmann, G., Rosenbaum, L., Jahn, A., Fechner, N., Zell, A. (2011). jcompoundmap-per: An open source java library and command-line tool for chemical fingerprints. Journal of Cheminformatics 3(1), 3

Huang, N., Shoichet, B.K., Irwin, J.J. (2006). Benchmarking sets for molecular docking. J Med Chem 49(23), 6789-6801

Huang, S.Y., Zou, X. (2010). Advances and challenges in protein-ligand docking. Int J Mol Sci 11(8), 3016-3034

Ivanciuc, O. (2007). Applications of Support Vector Machines. Chemistry Reviews in Computational Chemistry, Volume 23, 291-400.

Jorgensen, W.L., Duffy, E.M. (2002). Prediction of drug solubility from structure. Adv Drug Deliv Rev, 31;54(3):355-66.

Jorgensen, W. (2004). The many roles of computation in drug discovery. Science 303(5665), 1813-1818

Jorissen, R. N. and Gilson, M. K. (2005). Virtual Screening of Molecular Databases Using a Support Vector Machine. J. Chem Inf. Model., 45, 549–561

Kriegl, J. M., Arnhold, T., Beck, B., and Fox, T. (2005). Prediction of Human Cytochrome P450 Inhibition Using Support Vector Machines. QSAR Comb. Sci., 24, 491–502

Lee, D. E., Song, J. H., Song, S. O., and Yoon, E. S. (2005). Weighted Support Vector Machine for Quality Estimation in the Polymerization Process. Ind. Eng. Chem. Res., 44, 2101–2105

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953). Equation of state calculations by fast computing machines. J Chem Phys 21, 1087-1092

Morris, G., Goodsell, D., Halliday, R., Huey, R., Hart, W., Belew, R., Olson, A. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19(14), 1639-1662

NVIDIA. (2009). Whitepaper NVIDIA's Next Generation CUDA Compute Architecture: Fermi.

Pal, N.R., Panja, R. (2013). Finding short structural motifs for re-construction of proteins 3D structure. Applied Soft Computing, Volume 13, Issue 2, Pages 1214–1221

Peterson, K.L. (2007). Artificial Neural Networks and Their use in Chemistry. Reviews in Computational Chemistry, Volume 16.

Pérez-Sánchez, H., Wenzel, W. (2011). Optimization methods for virtual screening on novel computational architectures. Curr Comput Aided Drug Des 7(1), 2011,44-52

Romero Reyes, I.V., Fedyushkina, I.V., Skvortsov, V.S., Filimonov, D.A. (2013). Prediction of progesterone receptor inhibition by high-performance neural network algorithm. Internat. J. Math. Models and Methods Appl. Sci., 7(3) 303-310.

Sanchez, R., Sali, A (1998). Large-Scale Protein Structure Modeling of the Saccharomyces cerevisiae Genome. Proc Natl Acad Sci USA 95(23), 13597-13602

Sánchez-Linares, I., Pérez-Sánchez, H., Guerrero, G.D., Cecilia, J.M., García, J.M. (2011). Accelerating multiple target drug screening on gpus. In: Proceedings of the 9th International Conference on Computational Methods in Systems Biology (CMSB' 11), New York, NY, USA, ACM (2011) 95-102

Sánchez-Linares, I., Pérez-Sánchez, H., García, J.M. (2012a). Accelerating grid kernels for virtual screening on graphics processing units. In D'Hollander, E., Padua, D., eds.: Parallel Computing: Proceedings of the International Conference ParCo 2011. Volume 22., IOS, 413-420

Sánchez-Linares, I., Pérez-Sánchez, H., Cecilia, J., García, J. (2012b). High-throughput parallel blind virtual screening using bindsurf. BMC Bioinformatics 13(Suppl 14), S13

Schneider, G., Wrede, P. (1998). Artificial neural networks for computer-based molecular design. Progress in Biophysics and Molecular Biology 70(3), 175-222

Shoichet, B. K., Bodian, D. L., Kuntz, I. D. (1992). Molecular docking using shape descriptors. Journal of Computational Chemistry 13, 380–397.

Yao, L., Evans, J. A., and Rzhetsky, A. (2009). Novel opportunities for computational biology and sociology in drug discovery. Trends in Biotechnology 27, 531–540.

Taskinen, J., Yliruusi, J. (2003). Prediction of physicochemical properties based on neural network modelling. Advanced Drug Delivery Reviews Volume 55, Issue 9, 12, Pages 1163–1183.

Venables, W.N., Ripley, B.D. (2002). Modern Applied Statistics with S. Fourth edn. Springer, New York, ISBN 0-387-95457-0.

Wang, R., Lu, Y., Fang, X., Wang, S. (2004). An extensive test of 14 scoring functions using the PDBbind re_ned set of 800 protein-ligand complexes. J Chem Inform Comput Sci 44(6), 2114-2125

Warmuth, M.K., Liao, J. Rätsch, G., Mathieson, M., Putta, S., Lemmen, C. (2003). Active learning with support vector machines in the drug discovery process. J Chem Inf Comput Sci, 43(2):667-73.

Weisel, M., Kriegl, J.M., Schneider, G. (2010). Architectural repertoire of ligand-binding pockets on protein surfaces. Chembiochem,11(4):556-63.

Yuriev, E., Agostino, M., Ramsland, P.A. (2011). Challenges and advances in computational docking: 2009 in review. J Mol Recogn 24(2), 149-164